Hierarchical Clustering of Complex Symbolic Data and Application for Emitter Identification

Xu, Xin

2018-07

Xu , X , Lu , J & Wang , W 2018 , ' Hierarchical Clustering of Complex Symbolic Data and Application for Emitter Identification ' , Journal of Computer Science and Technology , vol. 33 , no. 4 , pp. 807-822 . https://doi.org/10.1007/s11390-018-1857-9

http://hdl.handle.net/10138/303906
https://doi.org/10.1007/s11390-018-1857-9

unspecified
acceptedVersion

Hierarchical Clustering of Complex Symbolic Data and Application for Emitter Identification

Xin Xu [1], Jiaheng Lu [2], Wei Wang [3], ACM, CCF member

[1] *Science and Technology on Information System Engineering Laboratory, Nanjing Research Institute of Electronic Engineering, Nanjing 210007, China*

[2] *Department of Computer Science, University of Helsinki, Helsinki 00014, Finland*

[3] *State Key Laboratory for Novel Software and Technology, Nanjing University, Nanjing 210046, China*

Email: flora.xin.xu@gmail.com; jiahenglu@gmail.com; ww@nju.edu.cn

Abstract It is well-known that the values of symbolic variables may take various forms such as an interval, a set of stochastic measurements of some underlying patterns or qualitative multi-values and so on. However, the majority of existing work in symbolic data analysis still focuses on interval values. Although some pioneering work in stochastic pattern based symbolic data and mixture of symbolic variables has been explored, it still lacks of flexibility and computation efficiency to make full use of the distinctive individual symbolic variables. Therefore, we bring forward a novel hierarchical clustering method with weighted general Jaccard distance and effective global pruning strategy for complex symbolic data and apply it to emitter identification. Extensive experiments indicate that our method has outperformed its peers in both computational efficiency and emitter identification accuracy.

Keywords symbolic data analysis, stochastic pattern, fuzzy interval, hierarchical clustering, emitter identification

---

This work was supported by National Natural Science Foundation of China (No. 61771177, 61701454) and Jiangsu Natural Science Foundation (No. BK20160147, BK20160148).

1 Introduction

In symbolic data analysis (SDA), the data complexity has gone beyond the classic data framework. Instead of possessing single values only, the symbolic variables usually appear in aggregate forms to represent certain homogeneous behaviours of objects. These aggregated variables have drawn more and more attention especially when it comes to the age of big data.

Generally, there are two categories of symbolic variables, quantitative and qualitative. The most common quantitative symbolic variable is the interval-valued one, where interval regions are provided. For example, studies show that most people within the age interval of [18, 45] are in favour of Military service. Meanwhile, the most common qualitative symbolic variable is the qualitative multi-valued one whose value is a finite subset of a category set with the corresponding weights, frequencies or probabilities to indicate how frequent or likely that category is for this element.

Recently, another type of quantitative symbolic data has become more and more popular, namely the stochastic pattern based symbolic data [1]. In the stochastic pattern based symbolic data, the variable values are sets of stochastic measurements. Examples of stochastic pattern based symbolic data objects include the aggregated behaviours of a customer group in online shopping, the daily heart rate measurements for a group of patients aged from 60 to 70, the parameter measurement sets of a certain type of radar emitters and so on. Here, each value of the stochastic pattern based symbolic variable is an instance of a stochastic pattern. Though some pioneering work in stochastic pattern based symbolic data has been conducted [2], it still suffers a high computational cost and lacks of robustness to various types of symbolic variables.

Nowadays, most SDA methods are restricted for the interval-valued symbolic data only. A considerable greater effort has been made for developing methods for interval-valued symbolic data. For instance, the representative SDA methods, including the univariate and bivariate descriptive statistics [3], factorial analysis [4], clustering [5], discriminant or unsupervised learning [6], linear regression [7] and time series analysis [8], are almost all designed for interval data.

As can be seen, existing SDA methods generally concentrate on one special type of symbolic data. In practical applications, there may be several different types of complex variables in the same symbolic data, either multi-valued or interval-valued or stochastic pattern based.

Table 1 illustrates a running example of complex symbolic data composed of a mixture

Table 1. An Example of Complex Symbolic Data

Observation	Heart Beat Rate (num. beat/min)	Blood Pressure (mmHg)	Appearance {height, hair color, skin color}	Class
o_1	{60, 69, 85, 100}	[62, 98]	{tall, black-hair, yellow-skin}	c_1
o_2	{61, 70, 84, 99}	[58, 102]	{tall, black-hair, yellow-skin}	c_1
o_3	{70, 86, 101}	[60, 100]	{tall, yellow-skin}	c_1
o_4	{71, 120}	[61, 101]	{tall, yellow-skin}	c_2
o_5	{69, 122}	[59, 99]	{tall, yellow-skin}	c_2
o_6	{70, 118}	[90, 120]	{yellow-skin}	c_3
o_7	{70, 120}	[93, 125]	{yellow-skin}	c_3

of qualitative multi-valued, interval-valued and stochastic pattern based variables. Specifically, attribute "heart beat rate" is stochastic pattern based, attribute "blood pressure" is interval-valued and attribute "appearance" is multi-valued. In such a case, none existing symbolic data analysis methods could be applied to discriminate the three different classes.

The benchmark interval data analysis methods are unable to discriminate class c_1 from class c_2, as the two classes are overlapping heavily on attribute "blood pressure". The stochastic pattern based methods are unable to discriminate class c_2 from c_3 either, since the two classes are overlapping on attribute "heart beat rate". However, all the three classes could be discriminated well when considering all the three attributes. Specifically, class c_3 is different from classes c_1 and c_2 on attribute "blood pressure"; class c_1 and c_2 are different on both the stochastic pattern based attribute "heart beat rate" and the multi-valued attribute "appearance".

In [9], a framework has been put forward to address complex symbolic data composed of a mixture of qualitative multi-valued, interval-valued and stochastic pattern based variables. It evaluates the similarity between a pair of symbolic variables for each data type separately and sums them up to produce a global similarity score. For example, for the running example in Table 1, it evaluates the similarity on symbolic variable "heart beat rate", "blood pressure" and "appearance" respectively and sums up the three similarity scores to get the global scores. Upon that, hierarchical clustering is applied and the symbolic data would be clustered into groups of interest (Table 7).

However, when it comes to real world application, e.g., emitter identification, it still lacks of flexibility and computation efficiency to make full use of the distinctive individual symbolic variables. Emitter identification is basically a classification task. Each training emitter observation is composed of a mixture of symbolic variable types and an emitter type. The task is to identify the emitter types given the complex symbolic observations. In this paper, we extend our previous approach by revisiting the similarity composition methods and evaluate it thoroughly in a real world emitter identification application.

Inspired by the above problems, we bring forward a novel hierarchical clustering method for complex symbolic pattern discovery and apply it to emitter identification. The major contributions are listed as follows:

- We propose the concept of weighted general Jaccard distance for flexible similarity evaluation on a pair of complex symbolic observations composed of interval-valued, multi-valued and stochastic pattern based variables;
- We develop a global pruning strategy for complex symbolic data to further enhance the computation efficiency;
- Extensive experiments on both synthetic and real-life emitter datasets have validated the efficiency and effectiveness of our method for application in emitter identification.

The rest of paper is organized as follows. We review related work in Section 2. Our hierarchical clustering method for complex symbolic data is formally presented in Section 3. In Section 4, we present the experimental results and apply our method to real-life emitter identification. The conclusion is made in Section 5.

2 Related Work

Our work belongs to symbolic data analysis (SDA). SDA was first introduced by E. Diday in the 1980s [1, 10, 11]. The aim of SDA is to address the need to represent and analyze the data which is unable to be represented in the classical data model. The pioneering SDA projects include two European research projects, "Symbolic Objects Data Analysis System" (SODAS) [12] and "Analysis System of Symbolic Official data" (ASSO) . The SODAS project was devoted for systematic development of data analysis methodologies for symbolic data and produced the first statistical package for SDA. Following the effort of SO-

ASSO, https://www.info.fundp.ac.be/asso/

DAS, the ASSO project continued to develop new SDA methodologies and expanded the statistical package. Meanwhile, the first book on SDA, "Analysis of Symbolic Data" [13] was formally published.

Generally, there are three typical types of symbolic variable, the qualitative multi-valued, interval-valued and stochastic pattern based. The symbolic variable is qualitative multi-valued if its values are finite subsets of the domain, interval-valued if an empirical distribution over a set of subintervals is given or stochastic pattern based if the variable values are sets of stochastic measurements corresponding to a certain stochastic process [1].

However, there has been quite a lot of effort in interval-valued symbolic data analysis. The benchmark SDA methods for interval-valued symbolic data analysis include the univariate and bivariate descriptive statistics [14], factorial analysis [4], clustering [5], discriminant or unsupervised learning [6], supervised learning [15], linear regression [7] and time series analysis [8]. Some of them have been adapted for histogram-valued data [8, 16]. And some fuzzy pattern mining approaches based on pre-defined interval structures have been explored [17] as well.

In term of similarity evaluation, our work is related with Jaccard index [18, 19]. The traditional Jaccard index [18], also known as the Jaccard similarity coefficient, is a statistic used for comparing the similarity and the diversity of sample sets. It measures similarity between finite sample sets, which is defined as the cardinality of the intersection of the two sample sets divided by the cardinality of the union of the two. A generalized Jaccard similarity [19] has been proposed to evaluate the similarity between two real-valued vectors of equal length. The Jaccard similarity coefficient is defined as the proportion of the sum of the minimum element values to that of the maximum element values. However, none of the variants is adaptable to the similarity evaluation of either interval-valued or stochastic pattern based symbolic variables yet.

In term of uncertainty processing strategy, our work is also related with the fuzzy pattern mining methods. Quite a large number of fuzzy pattern mining methods on uncertain data have been put forward to address the "fuzziness" in either item distribution [20, 21] or item specification [17, 22, 23]. On one hand, in order to cope with the fuzziness of item distribution, many probabilistic frequent item mining methods based on the probabilistic model have been put forward so that the frequentness probabilities of item sets could be approximated accurately [20, 21]. On the other hand, in order

to deal with the fuzziness of item specification, the fuzzy set theory [22, 23] and the interval structured approaches [17] have been applied as well. However, all these fuzzy pattern mining methods demand clear definitions of crystal item, fuzzy set or region specification, which is inappropriate in real applications.

Hierarchical clustering techniques [24, 25] have received quite much attention in various domains for partitioning objects into optimally homogeneous groups. The discovered clusters reflect certain empirically measured relations of similarity.

For multi-dimensional spatial data, various spatial query approaches [26, 27, 28] could be utilized to speed up the hierarchical clustering process. The closest pairs [26] in a spatial dataset could be identified efficiently with the branch-and-bound techniques [28] based on the R-tree index [27]. In such ways, the time complexity of hierarchical clustering on spatial datasets could be reduced to $\mathcal{O}(nlogn)$. However, these approaches are not applicable to our complex symbolic dataset since our similarity evaluation metric is different. Specifically, our general Jaccard index does not satisfy the $\delta-$inequality requirement of the spatial dataset claimed in [26].

Some efficient hierarchical clustering approaches for discrete datasets have been proposed as well. The pruning strategy in the "similarity join" approach [29] on records which are composed of token sets is rather similar to ours on qualitative multi-valued symbolic variables. It explores the prefix filtering, positional filtering and suffix filtering strategies for fast similarity evaluation based on the Jaccard similarity. However, it is restricted to the qualitative multi-valued variable and could not be applied to the stochastic pattern based one in our symbolic dataset. In addition, it is reported that the MapReduce strategy helps to speed up the hierarchical clustering significantly [30]. For instance, with the MapReduce framework, the top-k join approach [30] successfully reduces the time of web access log hierarchical clustering for user group discovery from 80 hours to 6 hours.

We also adopt hierarchical clustering for symbolic pattern discovery as the prior work [2, 9] did. In [2], a novel hierarchical clustering algorithm for stochastic pattern based symbolic data is proposed to conduct stochastic pattern discovery only. However, it is restricted for stochastic pattern only. Comparatively, besides the stochastic pattern, our method is available for qualitative multi-valued and interval-valued symbolic pattern discovery as well. And a framework has been put forward to address complex symbolic data composed of

a mixture of qualitative multi-valued, interval-valued and stochastic pattern based variables [9]. However, its pruning strategies are restricted for individual symbolic variables and it still lacks of a flexible general Jaccard distance calculation metric to make full use of the distinctive information from all the symbolic variables. As a result, it is not flexible enough for real applications yet. In this work, we bring forward a weighted general Jaccard distance calculation metric and a global pruning strategy to further enhance the robustness, flexibility and computation efficiency.

In this paper, instead of computing the general Jaccard distances for all the observation pairs on each symbolic variable, we calculate the the general Jaccard distance on multi-valued and stochastic pattern based symbolic variables with efficient similarity pruning first. The observation pairs below the similarity threshold are pruned away. Then, we further calculate the general Jaccard distance on interval-valued variables for the remaining observation pairs only.

3 Method

In this section, we formally propose our hierarchical clustering method for complex symbolic pattern discovery. Our method is composed of three major components: 1) similarity evaluation of symbolic variables, 2) distance matrix construction via similarity pruning and 3) symbolic pattern discovery via hierarchical clustering. The input of our method is the complex symbolic data consisted of a mixture of qualitative multi-valued, interval-valued and stochastic pattern based symbolic variables while the output is the set of discovered complex symbolic patterns. Table 2 summarizes the notations in our method.

Firstly, we propose a novel evaluation metric based on Jaccard index to evaluate the similarity for qualitative multi-valued, interval-valued and also the stochastic pattern based symbolic variables. Then, an effective pruning strategy is introduced to speed up the distance matrix construction process. And finally, a novel hierarchical clustering procedure [31] based on the general Jaccard index is outlined for the discovery of complex symbolic patterns composed of either qualitative multi-valued or interval-valued or stochastic pattern based symbolic variables.

The details of our hierarchical clustering method for complex symbolic pattern discovery are illustrated as follows.

Table 2. Notation

Symbol	Indication
C_i	Cluster candidate i
M_i	Value of the qualitative multi-valued symbolic variable in cluster candidate i
I_i	Value of the interval-valued symbolic variable in cluster candidate i
S_i	Value of the stochastic pattern based symbolic variable in cluster candidate i
S_{ir}	the $r-$th numeric measurement in stochastic numeric measurement set S_i
M_{ir}	the $r-$th discrete element in qualitative multi-valued set M_i
w_{ir}	Weight of the rth measurement/element in set S_i/M_i
I_{il}	Lower bound of interval region I_i
I_{iu}	Upper bound of interval region I_i
$MatchSet_M(M_i, M_j)$	Matched set between two qualitative multi-valued sets M_i and M_j
$MatchSet_I(I_i, I_j)$	Matched set between two interval regions I_i and I_j
$MatchSet_S(S_i, S_j)$	Matched set between two stochastic numeric measurement sets S_i and S_j
$Jaccard_M(M_i, M_j)$	General Jaccard index between two qualitative multi-valued sets M_i and M_j
$Jaccard_I(I_i, I_j)$	General Jaccard index between two interval regions I_i and I_j
$Jaccard_S(S_i, S_j)$	General Jaccard index between two stochastic numeric measurement sets S_i and S_j
$JaccardDist_M(.,.)$	General Jaccard distance between two qualitative multi-valued sets
$JaccardDist_I(.,.)$	General Jaccard distance between two interval regions
$JaccardDist_S(.,.)$	General Jaccard distance between two stochastic numeric measurement sets
$JaccardDist(.,.)$	General Jaccard distance between two symbolic observations
$MemSet_i$	Member set of cluster candidate i
Sup_i	Support of cluster candidate i
$Dis_{Single}(.,.)$	Distance between a pair of cluster candidates with the single linkage
δ	Approximation threshold
ϵ	Similarity threshold within range [0,1]
$minw$	Minimum weight threshold within range [0,1]
$minsup$	Minimum support threshold
Ω	Set of discovered complex symbolic patterns

3.1 Similarity Evaluation of Symbolic Variables

The traditional Jaccard index is only applicable for the qualitative multi-valued symbolic variable. Though the δ-Jaccard index has been put forward to evaluate the similarity between stochastic pattern based symbolic variables [2], it is not flexible enough for the interval-valued and the stochastic pattern based symbolic variables yet. For this reason, we propose a general Jaccard index for various symbolic variable types.

3.1.1 Matched Set

To evaluate the similarity between symbolic variables which are either qualitative multi-valued or interval-valued or stochastic pattern based, we first define the concept of matched set for the three different types of symbolic variables respectively.

The matched set between two qualitative multi-valued symbolic variables $M_i = \{M_{ip}\}_p$ and $M_j = \{M_{jq}\}_q$, denoted as $MatchSet_M(M_i, M_j)$, is defined as the set of common elements within the two sets, as shown in (1):

$$MatchSet_M(M_i, M_j) = M_i \bigcap M_j. \tag{1}$$

The matched set between two interval-valued symbolic variables $I_i = [I_{il}, I_{iu}]$ and $I_j = [I_{jl}, I_{ju}]$, $MatchSet_I(I_i, I_j)$, is calculated as their overlapping region, as illustrated in (2),

$$MatchSet_I(I_i, I_j) = \begin{cases} [\max(I_{il}, I_{jl}), \min(I_{iu}, I_{ju})], & if\ \max(I_{il}, I_{jl}) \le \min(I_{iu}, I_{ju}), \\ \emptyset, & otherwise. \end{cases} \tag{2}$$

As stated in [2], given a specified approximation threshold δ and a symmetric distance function $dist(x, y) = \frac{|x-y|}{\max(x,y)}$, the matched set between two stochastic numeric measurement sets $S_i = \{S_{ip}\}_p$ and $S_j = \{S_{jq}\}_q$ is defined as the set of their matched pairs within δ distance away, $MatchSet_S(S_i, S_j) = \{(S_{i1}, S_{jm_1}), (S_{i2}, S_{jm_2}), ..., , (S_{it}, S_{jm_t})\}$, where t is the number of matched pairs.

3.1.2 General Jaccard index

Based on the concept of matched set, we further propose the general Jaccard index for similarity evaluation of qualitative multi-valued, interval-valued and stochastic pattern based symbolic variables. The general Jaccard index between two qualitative multi-valued sets M_i and M_j is calculated as the proportion of the matched set size to the union size of M_i and M_j:

$$Jaccard_M(M_i, M_j) = \frac{|MatchSet_M(M_i, M_j)|}{|M_i| + |M_j| - |MatchSet_M(M_i, M_j)|}. \tag{3}$$

The general Jaccard index between two interval regions I_i and I_j is calculated as the proportion of matched interval length to the union interval

length:

$$Jaccard_I(I_i, I_j) = \frac{Len(MatchSet_I(I_i, I_j))}{Len(I_i) + Len(I_j) - Len(MatchSet_I(I_i, I_j))}, \quad (4)$$

where $Len()$ indicates the length of the corresponding interval region. The general Jaccard index between two stochastic numeric measurement sets S_i and S_j is calculated as the number of matched measurement pairs to the total number of distinct measurements after matching:

$$Jaccard_S(S_i, S_j) = \frac{|MatchSet_S(S_i, S_j)|}{|S_i| + |S_j| - |MatchSet_S(S_i, S_j)|}. \quad (5)$$

As can be observed, our general Jaccard indexes for all the three different types of symbolic variables vary between zero and one.

On the qualitative multi-valued attribute "appearance" in Table 1, observations o_1 and o_2 share three equal discrete values, thus they have a general Jaccard index of 1. On the interval-valued attribute "blood pressure", the general Jaccard index between observations o_1 and o_2 is $36/44 \approx 0.82$. For the stochastic pattern based attribute "heart beat rate", given the approximation threshold δ of value 0.1, observations o_1 and o_2 achieve a general Jaccard index of 1, since all their four pairs of stochastic numeric measurements are within δ distance away (Please refer to [2] for details).

3.2 Distance Matrix Construction via Similarity Pruning

Based on the proposed general Jaccard index, we construct the distance matrix for the complex symbolic observations via an effective pruning strategy.

We define the general Jaccard distance between two symbolic variables of a certain type as one minus the corresponding general Jaccard index, as shown in (6), (7) and (8):

$$JaccardDist_M(M_i, M_j) = 1 - Jaccard_M(M_i, M_j), \quad (6)$$

$$JaccardDist_I(I_i, I_j) = 1 - Jaccard_I(I_i, I_j), \quad (7)$$

$$JaccardDist_S(S_i, S_j) = 1 - Jaccard_S(S_i, S_j). \quad (8)$$

The general Jaccard distance between two symbolic observations is defined as the sum of weighted general Jaccard distances between the symbolic variables in the two observations, where α_A indicates the weight for symbolic variable A, as shown in (9):

$$JaccardDist(i, j) = \sum_A \alpha_A \times JaccardDist_A(A_i, A_j). \quad (9)$$

Inspired by the test statistic using pairwise similarity measures in [32], we extend it to our complex symbolic datasets. Given a set of complex symbolic observations with class labels, we define a modified test statistic d_A to evaluate the class discriminant power of each symbolic

variable A in the observation as the difference between the average within-class general Jaccard index and the average between-class general Jaccard index, as illustrated in (10):

$$d_A = \overline{Jaccard_A}_{\text{ within}} - \overline{Jaccard_A}_{\text{ between}}, \quad (10)$$

where $\overline{Jaccard_A}_{\text{ within}}$ indicates the average general Jaccard index on symbolic variable A for all pairs of observations from the same class and $\overline{Jaccard_A}_{\text{ between}}$ indicates the average general Jaccard index on symbolic variable A for all pairs of observations from different classes. Upon that, the the weight for each symbolic variable A could be inferred:

$$\alpha_A = \frac{d_A}{\sum_A d_A}. \quad (11)$$

As can be observed, the similarity evaluation and the distance calculation on the qualitative multi-valued and the stochastic pattern based symbolic attributes are the bottleneck. Therefore, we develop an effective pruning strategy to speed up the distance matrix construction process. The basic idea of our pruning strategy is to estimate the upper bound of the general Jaccard index of these symbolic variables and waive the distance calculation when the estimated upper bound is below the specified similarity threshold ϵ.

According to the definitions of general Jaccard index for qualitative multi-valued and stochastic pattern based symbolic variables, we can easily infer that the maximal general Jaccard index is achieved when the size of matched set is maximized. The formal rationale is provided in Lemma 1.

Lemma 1. *Suppose V_i and V_j are either two qualitative multi-valued sets or two stochastic numeric measurement sets whose sizes are $|V_i|$ and $|V_j|$ respectively, then the upper bound of general Jaccard index between sets V_i and V_j is*

$$upper_{\text{Jaccard}}(V_i, V_j) = \frac{\min(|V_i|, |V_j|)}{|V_i| + |V_j| - \min(|V_i|, |V_j|)}. \quad (12)$$

Proof. Since the maximum size of the matched set between sets V_i and V_j is $\min(|V_i|, |V_j|)$, the conclusion holds. □

Based on Lemma 1, we have designed a novel similarity pruning strategy for individual qualitative multi-valued and stochastic pattern based variables as follows: For each qualitative multi-valued or stochastic pattern based attribute, we rank the observations first in descending order of value set sizes and next in ascending order of original observation index. The larger size and smaller observation index is, the higher rank the observation would obtain.

Then, starting from the first observation in rank, we calculate the general Jaccard index for the qualitative multi-valued or the stochastic pattern based variable between the current observation and its successors in turn. Once

Table 3. Distance Matrix Construction on "Heart Beat Rate" via Similarity Pruning

$JaccardDist_S$	o_2	o_3	o_4	o_5	o_6	o_7
o_1	$\rightarrow 0.00$	$\rightarrow 0.25$	-	-	-	-
o_2		$\rightarrow 0.25$	-	-	-	-
o_3			-	-	-	-
o_4				$\rightarrow 0.00$	$\rightarrow 0.00$	$\rightarrow 0.00$
o_5					$\rightarrow 0.00$	$\rightarrow 0.00$
o_6						$\rightarrow 0.00$

Table 4. Distance Matrix Construction on "Appearance" via Similarity Pruning

$JaccardDist_M$	o_2	o_3	o_4	o_5	o_6	o_7
o_1	$\rightarrow 0.00$	$\rightarrow 0.33$	$\rightarrow 0.33$	$\rightarrow 0.33$	-	-
o_2		$\rightarrow 0.33$	$\rightarrow 0.33$	$\rightarrow 0.33$	-	-
o_3			$\rightarrow 0.00$	$\rightarrow 0.00$	-	-
o_4				$\rightarrow 0.00$	-	-
o_5					-	-
o_6						$\rightarrow 0.00$

the estimated upper bound of general Jaccard index is below the similarity threshold ϵ, the distance calculation for the current observation stops and starts the next round of calculation for the next observation in rank.

The calculation process could be safely pruned because once the estimated upper bound of general Jaccard index between the current observation o_i and its successor o_j is below the similarity threshold ϵ, the general Jaccard index between o_i and those successors ranked after o_j must be below threshold ϵ as well. The details of the rationale are given in Lemma 2.

Lemma 2. *Suppose attribute V in symbolic data D is either multi-valued or stochastic pattern based, ORD is the rank of the observations such that the observations are sorted first in descending order of value set sizes of V and next in ascending order of original observation index, ϵ is the specified similarity threshold. If the estimated upper bound of general Jaccard index between the current observation o_i and its successor o_j in rank ORD on attribute V is below ϵ, $upper_{\text{Jaccard}}(V_i, V_j) < \epsilon$, then for any value set of successor o_k of o_j, denoted as V_k,*

Table 5. Distance Matrix Construction on "Blood Pressure"

$JaccardDist_I$	o_2	o_3	o_4	o_5	o_6	o_7
o_1	0.18	0.10	-	-	-	-
o_2		0.09	-	-	-	-
o_3			-	-	-	-
o_4				0.10	-	-
o_5					-	-
o_6						0.23

Table 6. Distance Matrix Construction between Symbolic Observations

$JaccardDist$	o_2	o_3	o_4	o_5	o_6	o_7
o_1	0.06	0.23	-	-	-	-
o_2		0.22	-	-	-	-
o_3			-	-	-	-
o_4				0.03	-	-
o_5					-	-
o_6						0.08

we must have $upper_{\text{Jaccard}}(V_i, V_k) < \epsilon$.

Proof. Since $o_k \succ o_j$ in ORD rank, we have $|V_k| \leq |V_j|$. And since $upper_{\text{Jaccard}}(V_i, V_j) = \frac{|V_j|}{|V_i|} < \epsilon$, we have $upper_{\text{Jaccard}}(V_i, V_k) = \frac{|V_k|}{|V_i|} \leq upper_{\text{Jaccard}}(V_i, V_j) < \epsilon$. Therefore, the general Jaccard distance calculation between o_i and successors after o_j could be safely pruned. □

The observation pairs that do not satisfy the similarity threshold ϵ on either qualitative multi-valued or stochastic pattern based variables would be pruned. The corresponding distance calculation on the interval-valued attributes would be waived.

For instance, the rank of the seven observations in Table 1 on attribute "heart beat rate" is $o_1 \prec o_2 \prec o_3 \prec ...o_6 \prec o_7$. Given the similarity threshold $\epsilon = 0.6$, the distance calculation process would start from observation o_1. When it comes to successor o_4, the calculation process for observation o_1 stops, as the upper bound of general Jaccard index is below ϵ. Then the current observation will be updated to o_2 and the next round of calculation continues iteratively. Table 3, 4 and 5 illustrate

the process of distance matrix construction via similarity pruning first on attribute "heart beat rate", next on "appearance" and last on "blood pressure" respectively. Table 6 shows the process of distance matrix construction with equal weights of 1/3 between symbolic observations via a global similarity pruning on all the attributes. The units denoted with "-" in Tables 3, 4, 5 and 6 indicate the corresponding distance calculation has been pruned off.

Obviously, with the similarity pruning strategy conducted on all the symbolic variables simultaneously, a significant amount of computation cost could be saved.

3.3 Symbolic Pattern Discovery via Hierarchical Clustering

Upon the general Jaccard distance matrix, we discover the complex symbolic patterns via agglomerative hierarchical clustering of cluster candidates. For each cluster candidate C_i, its qualitative multi-valued set M_i is modelled as a set of weighted discrete elements, $M_i = \{M_{i1}, M_{i2}, ..., M_{i|M_i|}\}$. Similarly, its stochastic measurement set S_i is modelled as a set of weighted stochastic measurements, $S_i = \{S_{i1}, S_{i2}, ..., S_{i|S_i|}\}$. The values of these weights all vary within range [0, 1] to indicate the probability that the corresponding discrete element or stochastic measurement has a match in the current candidate cluster. The interval region I_i is modelled as $[I_{il}, I_{iu}]$. The member set of cluster candidate C_i is denoted as $MemSet_i$, indicating the set of symbolic observations it has covered. And the corresponding support value is the size of the member set, $Sup_i = |MemSet_i|$.

Subroutine SymbolicPatternDiscovery

Input Parameters:

- D: a complex symbolic dataset
- δ: the approximation threshold
- ϵ: the similarity threshold
- $minw$: the minimum weight threshold
- $minsup$: the minimum support threshold

Output:

- Ω: the set of discovered complex symbolic patterns

1. **for** each pair of observation o_i and $o_j \in D$ that $i < j$ **do**
2. flag[i,j]=true
3. **for** each multi-valued or stochastic pattern based attribute V **do**
4. set VS and Ord order; $cur = 1$;
5. **while** $cur < |VS|$ **do**
6. $suc = cur + 1$
7. **while** $!flag[Ord[cur], Ord[suc]]$ and $suc < |VS|$ **do**
8. $suc = suc + 1$
9. **while** $suc \leq |VS|$ and $\frac{|VS[Ord[suc]]|}{|VS[Ord[cur]]|} \geq \epsilon$ **do**
10. $index = Jaccard_V(VS[Ord[cur]], VS[Ord[suc]])$
11. **if** index$< \epsilon$ **then**
12. flag[Ord[cur], Ord[suc]]=false
13. **else**
14. $JaccardDist_V(Ord[cur], Ord[suc])$=1-index
15. $suc = suc + 1$
16. $cur = cur + 1$
17. **for** each interval-valued attribute I **do**
18. **for** each pair of observations o_i and $o_j \in D$ and $i < j$ and flag[i,j]=true **do**
19. index=$Jaccard_I(i, j)$
20. **if** index$< \epsilon$ **then**
21. flag[i, j]=false
22. **else**
23. $JaccardDist_I(i, j) = 1 - index$
24. **for** each pair of observations i and j **do**
25. **if** $flag[i, j] = TRUE$ **then**
26. $JaccardDist[i, j] = \sum_{A \in \{I,M,S\}} JaccardDist_A[i, j]$
27. **else**
28. $JaccardDist[i, j] = 1.0$
29. $\Omega = \emptyset$
30. **for** cluster candidate C_i **do**
31. $C_i = o_i$; $MemSet_i = \{o_i\}$; $Sup_i = 1$;
32. $CS = \{C_i\}$
33. **repeat** find closest $C_x, C_y \in CS$ below $1 - \epsilon$ **do**
34. merge C_x and C_y into $C_{x'}$; update symbolic variable models;
35. $MemSet_{x'} = MemSet_x \cup MemSet_y$; $Sup_{x'} = Sup_x + Sup_y$;
36. output the set of symbolic patterns Ω above $minsup$.

Fig. 1. Subroutine of symbolic pattern discovery.

Each cluster candidate is initialized with

an individual symbolic observation from symbolic dataset D. Then, the cluster candidates would merge with one another agglomeratively as long as the general Jaccard indexes between them are above the specified similarity threshold ϵ in each attribute dimension. During the above hierarchical clustering, the qualitative multi-valued sets, interval regions, stochastic measurement sets, member sets and supports of the cluster candidates would be updated dynamically all along the way. Also, a minimum weight threshold *minw* is applied so that the qualitative elements and stochastic measurements below threshold *minw* would be removed from the models.

The set of complex stochastic patterns, denoted as Ω, would be discovered from the final cluster candidates above the minimum support threshold *minsup*. The details of the complex symbolic pattern discovery subroutine is illustrated in Fig. 1.

Cluster Candidate Initialization

Each cluster candidate C_i is initialized with an individual symbolic observation. Specifically, for the qualitative multi-valued set and the stochastic numeric measurement set, the weights of the corresponding elements and measurements are all initialized as 1. The member set $MemSet_i$ is initialized as the corresponding symbolic observation o_i, $MemSet_i = \{o_i\}$, and the support Sup_i is initialized as 1, $Sup_i = 1$.

For the running example in Table 1, a cluster candidate C_1 could be initialized with o_1 such that $M_1 = \{$tall, black-hair, yellow-skin$\}$, $I_1 = [62, 98]$ and $S_1 = \{60, 69, 85, 100\}$. The element weights of M_1 and S_1 are all initialized as 1. For instance, w_{11} of M_1 indicates the weight of "tall" element which is initialized as 1. Likewise, cluster candidates C_2 and C_3 could be initialized with o_2 and o_3 respectively. For cluster candidate C_2, we have $M_2 = \{$tall, black-hair, yellow-skin$\}$, $I_2 = [58, 102]$ and $S_2 = \{61, 70, 84, 99\}$. And for cluster candidate C_3, we have $M_3 = \{$tall, yellow-skin$\}$, $I_3 = [60, 100]$ and $S_3 = \{70, 86, 101\}$.The supports of these cluster candidates are all initialized as 1, $Sup_1 = Sup_2 = Sup_3 = 1$.

Cluster Candidate Update

In this work, we make use of the single-linkage scheme during agglomerative hierarchical clustering. The general Jaccard distance between two cluster candidates, C_x and C_y, is defined as the minimum general Jaccard distance between the members from the two cluster candidates, as shown in (13):

$$dist_{\text{Single}}(C_x, C_y) = \min_{o_i \in MemSet_x, o_j \in MemSet_y} JaccardDist(i, j). \tag{13}$$

Of course, besides the single linkage, complete linkage and average linkage could be applied as well.

During the process of hierarchical clustering, the pair of cluster candidates $(C_x,\ C_y)$ with the minimum general Jaccard distance below threshold $1-\epsilon$ on each symbolic attribute would merge into a new cluster candidate $C_{x'}$. Specifically, the agglomerative merging of stochastic pattern based attribute proceeds just as that in [2]. For the qualitative multi-valued attribute, the agglomerative merging process is similar.

Firstly, the matched set $MatchSet(M_x,\ M_y)$ between the pair of qualitative multi-valued sets M_x and M_y is inferred. Then, for each matched element $M_{xp_k} = M_{yq_k} \in MatchSet_M(M_x,\ M_y),\ 1 \leq k \leq |MatchSet_M(M_x,\ M_y)|$, the associated element weight w_k would be generated according to (14) :

$$w_k = \frac{w_{xp_k} \times Sup_x + w_{yq_k} \times Sup_y}{Sup_x + Sup_y}. \tag{14}$$

As indicated in Table 6, cluster candidate C_1 would merge with candidate C_2 into cluster candidate C_1'. The matched set of their qualitative multi-valued sets $MatchSet(M_1,\ M_2)$ is {tall, black-hair, yellow-skin}, where the number of matched elements is three and the associate weights are all updated to one according to (14). The qualitative multi-valued set of cluster candidate C_1', denoted as M_1', thus becomes {tall, black-hair, yellow-skin} and the support of cluster candidate C_1' is updated to 2.

And for each unmatched element, either $M_{xp'}$ from set M_x or $M_{yq'}$ from set M_y, the corresponding weight w_r in the merged cluster candidate $C_{x'}$ would be calculated as shown in (15):

$$w_r = \begin{cases} \frac{w_{xp'} \times Sup_x}{Sup_x + Sup_y} & M_{xp'} \in M_x \ and \ M_{xp'} \ is \ unmatched \\ \frac{w_{yq'} \times Sup_y}{Sup_x + Sup_y} & M_{yq'} \in M_y \ and \ M_{yq'} \ is \ unmatched \end{cases}. \tag{15}$$

For instance, when cluster candidate C_1' further merges with candidate C_3, the element "black-hair" in $M_1' =$ {tall, black-hair, yellow-skin} has no match in $M_3 =$ {tall, yellow-skin}. According to (15), the weight of element "black-hair" is updated to 2/3, as its original weights are 1 for both C_1' and C_3 and its original supports are 2 and 1 for C_1' and C_3 respectively.

Similar to the update of stochastic patterns, we generally keep the discrete elements whose weights are above threshold $minw$. The elements with weights below threshold $minw$ are considered as noises and thus are pruned.

For the interval-valued attribute, given an interval $I_x = [I_{xl},\ I_{xu}]$ from cluster candidate C_x and the interval region $I_y = [I_{yl},\ I_{yu}]$ from cluster candidate C_y, the interval lower bound and upper bound for the merged cluster candidate $C_{x'}$ would be updated as well, as shown in (16) and (17) respectively:

Table 7. Three Discovered Complex Symbolic Patterns

Pattern	Heart Beat Rate	Blood Pressure	Appearance	Member Set
P_1	{60.5, 70, 85, 100}	[60, 100]	{tall, black-hair, yellow-skin}	$\{o_1, o_2, o_3\}$
P_2	{70, 121}	[60, 100]	{tall, yellow-skin}	$\{o_4, o_5\}$
P_3	{70, 119}	[91, 5, 122.5]	{yellow-skin}	$\{o_6, o_7\}$

$$I_{x'l} = \frac{I_{xl} \times Sup_x + I_{yl} \times Sup_y}{Sup_x + Sup_y}, \tag{16}$$

$$I_{x'u} = \frac{I_{xu} \times Sup_x + I_{yu} \times Sup_y}{Sup_x + Sup_y}. \tag{17}$$

Meanwhile, the support and member set of the new cluster candidate $C_{x'}$ would be calculated as well, as shown in (18) and (19):

$$Sup_{x'} = Sup_x + Sup_y, \tag{18}$$

$$MemSet_{x'} = MemSet_x \cup MemSet_y. \tag{19}$$

Global Similarity Pruning Strategy

Note that the distance matrix construction process typically starts from the qualitative multi-valued or stochastic pattern based symbolic variables and ends with the interval-valued ones. In addition, with our global similarity pruning strategy, once we find that the pairs of observations whose Jaccard distances on the current symbolic attributes do not satisfy the ϵ threshold, the corresponding distance calculation on other symbolic attributes would be waived. This pruning strategy ensures that the cluster candidates merge with each other only when they satisfy the similarity threshold ϵ on all the attributes. Finally, the complex stochastic patterns satisfying threshold $minsup$ would be discovered after hierarchical merging.

For instance, given the approximation threshold $\delta = 0.1$, similarity threshold $\epsilon = 0.6$, minimum weight threshold $minw = 0.5$ and minimum support threshold $minsup = 2$, the final distance matrix constructed for the running example in Table 1 is illustrated in Table 6. The units marked with "-" indicate the corresponding general Jaccard distance calculation has been skipped. Three complex symbolic patterns are discovered as illustrated in Table 7, which are representative for class c_1, c_2 and c_3 respectively.

4 Results

We evaluated our hierarchical clustering method for complex symbolic pattern discovery on a series of synthetic datasets and applied it for real-life emitter identification. Experiments were conducted on a Dell PC running Microsoft Windows XP with a Pentium dual-core CPU of

2.6GHz and a 4G RAM.

The synthetic datasets are composed of three types of symbolic attributes, the qualitative multi-valued, the interval-valued and the stochastic pattern based. For the qualitative multi-valued symbolic attribute, six qualitative multi-valued sets of different lengths varying from 3 to 8 are embedded. For an interval-valued attribute, seven interval regions are embedded. And for the stochastic pattern based attribute, three overlapping stochastic patterns of length 3, 5 and 8 are embedded respectively.

The stochastic numeric measurements and the interval bound values all comply with a normal distribution $\mathcal{N}orm(p,\ sd)$, where p is the underlying true value, $sd = c \times p$ is the standard deviation and the coefficient c is varied between 0.1 and 0.5.

To evaluate the robustness of our method to value missing, a missing probability *mprob* was applied and set as 20% in default. We made use of a data generator with a random variable R for missing measurement simulation. The values of variable R follow a uniform distribution in the range of [0, 1]. In case variable R is below *mprob*, the corresponding discrete element in the qualitative multi-valued set and the measurement in the stochastic pattern would be missed.

The real-life airborne emitter parameter dataset consists of 7k symbolic observations. Each observation consists of one qualitative multi-valued "working mode" parameter, one interval-valued "RF" (radio frequency) parameter, one stochastic pattern based "PRI" (pulse repetition interval) parameter and a class label indicating the emitter type, as shown in Table 8. There are three different emitter types, denoted as C_1, C_2 and C_3 respectively. In addition, an independent test dataset is provided to validate the discovered complex symbolic patterns.

Specifically, the "PRI measurement" attribute value is a set of stochastic measurements of pulse repetitive interval for the emitter. The "RF interval" attribute value is an interval composed of the lower and upper bound of the radio frequency measurements of the emitter. The "working mode" attribute is composed of a set of discrete values describing the emitter working mode. Particularly, the "working mode" attribute is composed of a set of discrete values : Air (the emitter platform is an airplane, etc.), Ground (the emitter platform is a stationary one on the ground), Sea (the emitter platform is a ship, etc.), RF_low (the radio frequency measurements are in the low region), RF_mid (the radio frequency measurements are in the middle region), RF_high (the radio frequency measurements are in the high region),

Table 8. The Structure of the Real-life Airborne Emitter Parameter Dataset

Observation	PRI Measurements	RF Interval	Working Mode	Emitter Type
index	{measurement1, measurement2, ...}	[RF lower bound, RF upper bound]	{Air, Ground, Sea, ...}	C_1, C_2 or C_3

PRI_low (the PRI measurements are in the low region), PRI_mid (the PRI measurements are in the middle region), PRI_high (the PRI measurements are in the high region), Pulse_group (the working mode of the emitter PRI parameter).

We validated the efficiency and effectiveness of our hierarchical clustering method on a large number of synthetic datasets. In term of efficiency evaluation, we examined the usefulness of our similarity pruning strategy for general Jaccard distance calculation and symbolic pattern discovery, and tested the scalability of our method by varying the number of attributes. In term of effectiveness evaluation, we compared the discovered stochastic patterns against the underlying true ones in term of general Jaccard index. To evaluate the potential usefulness of our method in real applications, we also applied our hierarchical clustering method for complex symbolic pattern discovery in emitter identification.

In the default setting, we fixed the approximation threshold δ as 0.1 for the stochastic pattern based attributes. We also set the similarity threshold ϵ as 0.8 and the minimum weight threshold $minw$ as 0.5 for both the qualitative multi-valued and stochastic pattern based attributes. The minimum support threshold $minsup$ was fixed as 0.1.

4.1 Efficiency Evaluation

To evaluate the efficiency of our method, we compared the computational time (in seconds) of general Jaccard distance calculation and symbolic pattern discovery when varying the similarity threshold ϵ. We also examined the scalability of our method when varying the number of attributes.

4.1.1 Similarity Pruning for Distance Calculation

During the experiments, we evaluated the similarity pruning strategy on both the qualitative multi-valued and the stochastic pattern based symbolic attributes.

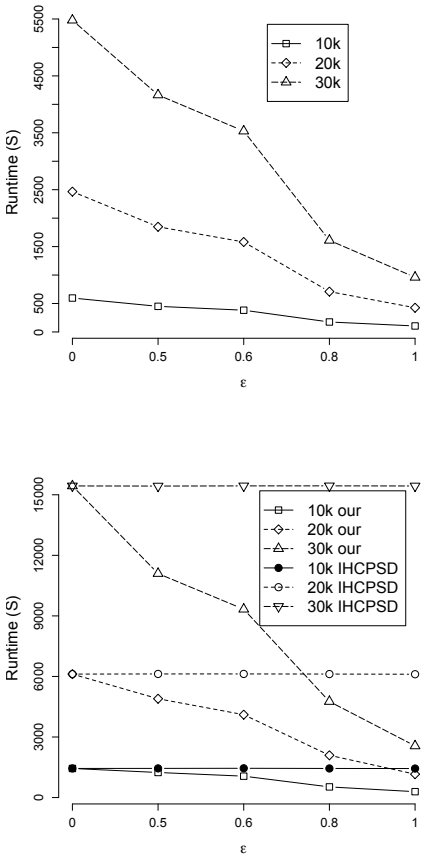


Fig. 2. Computational time of distance calculation when varying ϵ. (a) Qualitative multi-valued attribute. (b) Stochastic pattern based attribute.

As clarified in Lemma 2, the sizes of a pair of qualitative multi-valued variables or a pair stochastic measurement sets must be at least epsilon of each other to satisfy the similarity constraint. With the increase of the similarity threshold ϵ, the number of qualified multi-valued and stochastic measurement set pairs decreased significantly, and thus the amount of Jaccard distance calculation reduced. As a result, we can see a significant decrease in the computation time with the rising of threshold ϵ. Specifically, when the threshold ϵ was increased from 0 to 1, the computation time on the qualitative multi-valued attribute was decreased from around 600 seconds to 100 seconds with the data size 10k, decreased from around 2500 seconds to 400 seconds with the data size 20k and from around 5500 seconds to 950 seconds with the data size 30k, as shown in Fig. 2(a). Meanwhile, when threshold ϵ increased from 0.0 to 1.0, the computation time on the stochastic pattern based attribute was decreased from around 1440 seconds to 290 seconds with the data size 10k, decreased from around 6100 seconds to 1150 seconds with the data size 20k and from around 15400 seconds to 2600 seconds with the data size 30k, as shown in Fig. 2(b).

For the stochastic pattern based method IHCPSD (Incremental Hierarchical Clustering algorithm for stochastic Pattern-based Symbolic Data) [2], all pairs of stochastic measurement sets have to be compared. Therefore, our method has outperformed the IHCPSD method significantly in term of efficiency.

4.1.2 Scalability in Distance Calculation

In scalability evaluation, we compared the computational time of general Jaccard distance calculation on qualitative multi-valued,

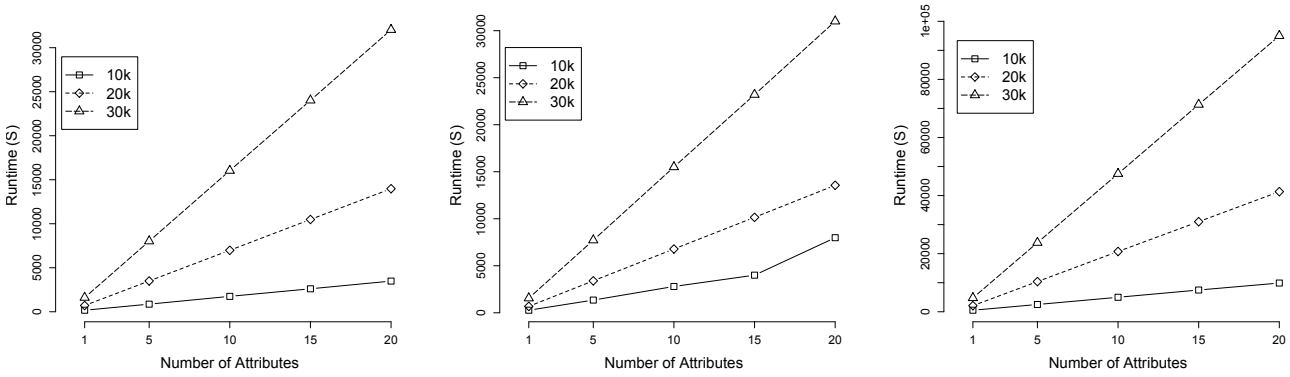


Fig. 3. Computational time of distance calculation when varying the number of attributes. (a) Qualitative multi-valued attribute. (b) Interval-valued attribute. (c) Stochastic pattern based attribute.

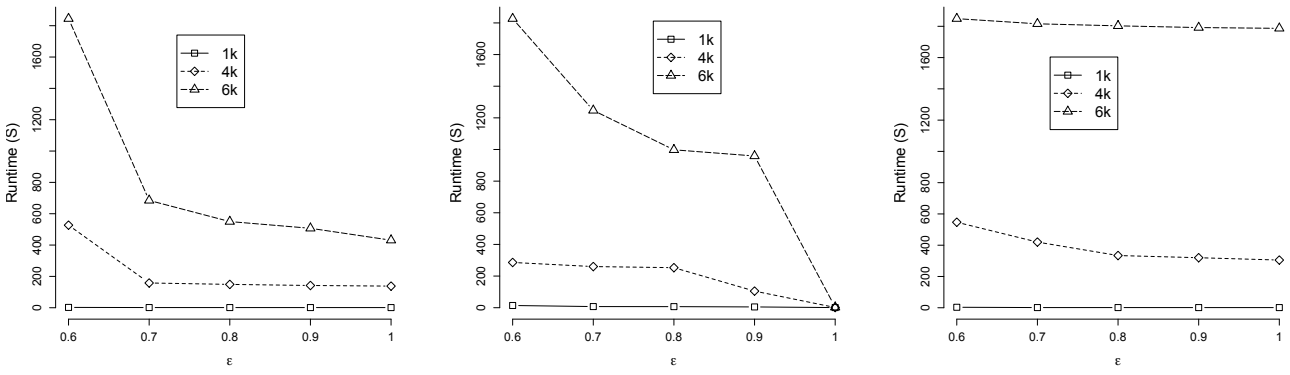


Fig. 4. Computational time of pattern discovery when varying ϵ. (a) Qualitative multi-valued attribute. (b) Interval-valued attribute. (c) Stochastic pattern based attribute

interval-valued and stochastic pattern based symbolic attributes respectively when varying the number of attributes. We presented the experimental results in Figs. 3(a), 3(b) and 3(c) respectively. As the computational cost of general Jaccard distance calculation for all the three types of symbolic attributes was approximately linear w.r.t. the number of attributes, the computational time increased approximately linearly with the increase of the number of attributes.

Generally, the computational time of general Jaccard distance calculation on the interval-valued attribute was the lowest and that on the stochastic pattern based attribute was the highest. However, as there was no similarity pruning for the interval-valued attributes, we observed a longer computational time for the synthetic dataset of size 10k on the interval-valued attribute than that on the qualitative multi-valued attribute.

4.1.3 Similarity Pruning for Pattern Discovery

In addition, we evaluated the computational time of pattern discovery via similarity pruning. We varied the similarity threshold ϵ from 0.6 to 1.0 and compared the corresponding runtime on the synthetic datasets.

The higher the similarity threshold was, the fewer cluster candidates merged, and viceversa. As a result, the computational time was negatively correlated with similarity threshold ϵ. As can be seen from Fig. 4, with the increase of ϵ, the runtime of pattern discovery decreased significantly, especially for the interval-valued symbolic variables. This also indicated that the interval-valued symbolic pattern discovery was more sensitive to the similarity threshold.

4.2 Effectiveness Evaluation

To evaluate the effectiveness of discovered complex symbolic patterns, we simulated the noises by varying the missing probability parameter *mprob* between 0.1 and 0.5 for both the qualitative multi-valued and stochastic pattern based symbolic attributes and varying the coefficient c between 0.1 and 0.5 for the interval-valued symbolic attributes.

When parameter *mprob* was set as 0.5, there was a probability of 50% that the corresponding qualitative multi-valued element and stochastic numeric measurement would be missed during the data simulation. When parameter *mprob* was set as 0.1, the probability of missing was 10%. Likewise, the larger the value of coefficient c is, the larger noises the synthetic data would have.

Firstly, we evaluated the effectiveness of discovered symbolic patterns on individual

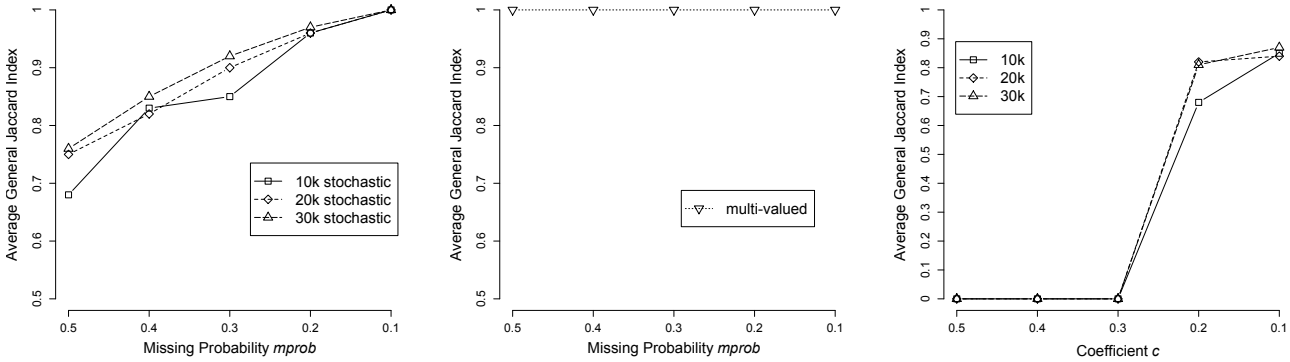


Fig. 5. Evaluation of effectiveness of discovered symbolic patterns on individual symbolic variables. (a) Stochastic pattern. (b) Multi-valued pattern. (c) Interval pattern.

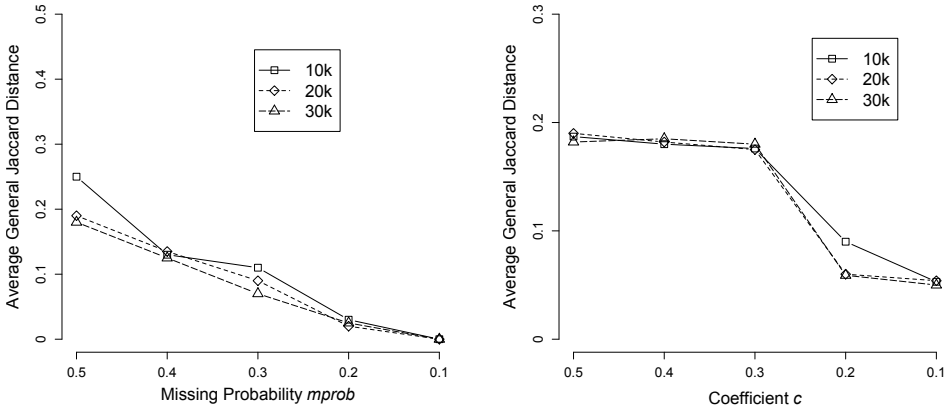


Fig. 6. Evaluation of effectiveness of discovered symbolic patterns on complex symbolic observations. (a) Average general Jaccard distance when varying the missing probability. (b) Average general Jaccard distance when varying parameter c.

symbolic variables. Given the discovered symbolic pattern on each individual symbolic variable, we assigned it to the closest underlying true ones and calculated the general Jaccard index between them. In this way, we could obtain the average general Jaccard indexes for the stochastic pattern, qualitative multi-valued pattern and interval pattern, as illustrated in Figs 5(a), 5(b) and 5(c) respectively. The higher the average general Jaccard index was, the more accurate the discovered symbolic patterns were.

As illustrated in Figs. 5(a) and 5(b), when the missing probability parameter *mprob* was varied from 0.5 to 0.1, all the qualitative multi-valued symbolic patterns were discovered successfully with an average general Jaccard index value of 1. As for the stochastic pattern based symbolic variable, the average general Jaccard index between the discovered stochastic patterns and the associated true ones was around 0.7 when parameter *mprob* was 0.5. This indicated that around 70% stochastic numeric measurements in the stochastic pattern have been discovered on average. When parameter *mprob* was either 0.4 or 0.3, more than 80% stochastic numeric measurements in the stochastic pattern have been discovered. And when parameter *mprob* was 0.1, all the stochastic measurements in the stochastic pattern have been discovered.

The experimental results for the interval patterns were illustrated in Fig. 5(c). When coefficient c was varied from 0.5 to 0.1, the mean values of the general Jaccard indexes for the interval-valued symbolic variable increased significantly. When coefficient c was above or equal to 0.3, none of the underlying true interval patterns were discovered and thus the resulted mean values of the calculated general Jaccard indexes were zeros. When coefficient c was 0.2, approximately 70% to 80% of the underlying true interval regions were successfully discovered. And when coefficient c was 0.1, around 85% of the underlying true interval regions were successfully discovered.

Next, we evaluated the effectiveness of discovered symbolic patterns on complex symbolic observations. We simulated the complex symbolic datasets composed of a qualitative multi-valued, a stochastic pattern based and an interval-valued variable. We calculated the average general distance between the discovered symbolic patterns on all the three symbolic variables and the underlying true ones. As can be seen in Figs. 6(a) and 6(b), with the decrease in missing probability and parameter c, the average general distances decrease significantly.

As can be seen, our method was robust to

Table 9. Emitter Identification Accuracy on the Transformed Data with Varying Variable Weights v.s. on the Mean & Range Data

Accuracy (%)	Mean & Range	Pattern-based Data Transformation				
		(0.33, 0.33, 0.33)	(1, 0, 0)	(0, 1, 0)	(0, 0, 1)	(0.1, 0.15, 0.75)
Naive Bayes	78.7	87.2	65.5	69.2	86.6	**89.5**
Logistic Regression	82.9	88.5	66.4	70.5	86.5	**90.2**
Multilayer Perceptron	90	95.2	68.2	78.6	94.3	**96.3**
RBFNetwork	88.6	93.2	65.5	68.6	93.4	**94.2**
SVM	78.7	85.3	62.8	67.5	84.4	**88.4**
KNN	92.3	94.2	67.5	71	94.2	**96.1**
Decision Tree	92.8	93.8	66.3	73.5	94.2	**95.2**

the noises and missing values in the qualitative multi-valued, the interval-valued and the stochastic pattern based symbolic attributes.

4.3 Application for Emitter Identification

Firstly, we discovered the complex symbolic patterns with the approximation threshold $\delta = 0.05$, the similarity threshold $\epsilon = 0.5$, the minimum weight threshold $minw = 0.2$, the minimum support threshold $minsup = 0.05\times$ the training dateset size.

Then, we applied the discovered complex symbolic patterns for emitter data transformation. Specifically, we selected the set of top discriminating symbolic patterns Ω and calculated the general Jaccard distance values between the discriminating patterns and the observations as provided in (9). In this way, each observation is transformed into a set of general Jaccard distance values, one for each discriminating symbolic pattern [2]. And the whole original complex symbolic dataset would be transformed into the classical data format with the corresponding general Jaccard distance values.

The class discriminating power of each discovered complex symbolic pattern P_i in Ω is evaluated by its pattern confidence $patconf_i$. The $patconf_i$ value is calculated as the maximum class distribution of the corresponding member set $MemSet_i$, as shown in (20):

$$patconf_i = \frac{\max_c |MemSet_i[c]|}{|MemSet_i|}, \tag{20}$$

where c indicates a certain emitter type whose class distribution rate in the member set $MemSei_i$ is the maximal one among all the emitter types.

For example, suppose the member set of a complex symbolic pattern q was composed

of 150 members, 100 from emitter type C_1, 20 from emitter type C_2 and 30 from emitter type C_3. Then the corresponding pattern confidence would be 0.67, as the maximum class distribution was obtained in emitter type C_1 whose distribution rate in the member set is 0.67.

We ranked the discovered complex symbolic patterns in descending order of pattern confidence values. The top twenty discriminating complex symbolic patterns from Ω with the highest pattern confidence values were then selected for data transformation.

With each selected complex symbolic pattern, the original symbolic observation would be transformed into a general Jaccard distance. In this way, the original complex symbolic dataset could be transformed into one composed of twenty columns. After the transformation, the classical data analysis approaches could be applied straightforward.

Finally, we compared the emitter type identification accuracy on the pattern-transformed emitter parameter data against that on the corresponding "Mean & Range" dataset. In the "Mean & Range" dataset, the stochastic pattern based value sets and interval regions were simply converted to the mean and range of the corresponding measurements. During experiments, we applied seven classification methods, the benchmark Naive Bayes, Logistic Regression, Multilayer Perceptron, RBFNetwork, SVM, KNN and Decision Tree. Please note that the IHCPSD method [2] was unable to deal with the interval regions and the qualitative multi-value sets for the above pattern-based data transformation.

Table 9 illustrates the emitter identification accuracy of seven benchmark classification methods on the transformed emitter data with varying variable weights against that on the mean & range emitter data. The highest accuracy achieved is highlighted in bold. With our method, a weight vector of (0.1, 0.15, 0.75) was assigned for the multi-valued, interval-valued and stochastic pattern based variables respectively according to (11). Alternatively, a weigh vector of (0.33, 0.33, 0.33) was assigned for the equal-weight approach in [9] . In addition, the weight vectors of (1, 0, 0), (0, 1, 0) and (0, 0, 1) were used for a single one symbolic variable. We set the weight for one symbolic variable as 1 and the remaining ones as 0 to obtain the "multi-valued only", "interval only" and "stochastic pattern only" results respectively. As can be observed, our method outperformed both the equal-weight approach and the single-one-variable approaches on the transformed dataset. The identification accuracy of our method on the transformed dataset is also higher than that on the mean & range

dataset. This is because we have made a better use of the underlying complex symbolic variables with our flexible weighted general Jaccard distance.

5 Conclusion

In this paper, we proposed a novel hierarchical clustering method for complex symbolic pattern discovery. To our knowledge, this is the first algorithm that not only deals with complex symbolic data of various types but also is adaptable for application in emitter identification. Experimental results indicate that it is robust to missing values and noises and it outperforms the peers in term of both efficiency and effectiveness.

References

[1] Noirhomme-Fraiture M, Brito P. Far beyond the classical data models: Symbolic data analysis. *Statistical Analysis and Data Mining: the ASA Data Science Journal*, 2011, 4(2): pp.157-170.

[2] Xu X, Lu J H, Wang W. Incremental hierarchical clustering of stochastic pattern based symbolic data. In *Advances in Knowledge Discovery and Data Mining*, Bailey J, Khan L, Washio T et al. (eds.), Springer, 2016, pp.156-167.

[3] Yu X C, He H, Hu D, Zhou W. Land cover classification of remote sensing imagery based on interval-valued data fuzzy c-means algorithm. *Science China*, 2014, 57(6): pp.1306-1313.

[4] Lauro C, Verde R, Irpino A. Generalized canonical analysis, In *Symbolic Data Analysis and the Sodas Software*, Diday E and Noirhomme-Fraiture M (eds.), Chichester, Wiley, 2008, pp.313-330.

[5] de A. T. de Carvalho F, M C R de Souza R. Unsupervised pattern recognition models for mixed feature-type symbolic data. *Pattern Recognition Letter*, 2010, 31(5): pp430-443.

[6] Rasson J P, Pircon J Y, Lallemand P, Adans S. Unsupervised divisive classification, In *Symbolic Data Analysis and the Sodas Software*, Diday E and Noirhomme-Fraiture M (eds.), Chichester, Wiley, 2008, pp.149-156.

[7] Neto L and de A. T. de Carvalho F. Constrained linear regression models for symbolic interval-valued variables. *Computational Statistics & Data Analysis*, 2010, 54(2): pp.333-347.

[8] Arroyo J, González-Rivera G, Maté C. Forecasting with interval and histogram

data. Some financial applications, In *Handbook of Empirical Economics and Finance*, Ullah A, Giles D, Balakrishnan N, Schucany W, Schilling E (eds.), Chapman and Hall/CRC, New York, 2010.

[9] Xu X. A novel hierarchical clustering framework for complex symbolic data exploration. In *Proc. the 32nd IEEE International Conference on Data Engineering Workshops*, 2016.

[10] Diday E. The symbolic approach in clustering and related methods of data analysis: the basic choices. In *Classification and Related Methods of Data Analysis, Proc. IFCS'87*, Bock H H (ed.), Aachen, North Holland, Amsterdam, 1988, pp673-684.

[11] Diday E. Introduction $\grave{a}$ l$'$ approche symbolique en analyse des données, *RAIRO-Operations Research*, 1989, 23(2): pp.193-236.

[12] Diday E, Noirhomme-Fraiture M. *Symbolic data analysis and the SODAS Software*. Wiley-Interscience New York, 2008.

[13] Bock H H, Diday E. Analysis of symbolic data, exploratory methods for extracting statistical information from complex data. Berlin-Heidelberg, Springer-Verlag, 2000.

[14] Billard L. Sample covariance functions for complex quantitative data, In *Proc. the Joint Meeting of 4th World Conference of the IASC and 6th Conference of the Asian Regional Section of the IASC on Computational Statistics & Data Analysis*, Yokohama, Japan, 2008.

[15] Lin C M, Chen Y M, Hsueh C S. A self-organizing interval type-2 fuzzy neural network for radar emitter identification. *International Journal of Fuzzy Systems*, 2014, 16(1).

[16] González-Rivera G, Arroyo J. Time series modeling of histogram-valued data: The daily histogram time series of S&P500 intradaily returns. *Int. J. Forecasting*, 2012, 28(1): pp.20-33.

[17] Mehdi K, Sergei O K, Amedeo N. Revisiting numerical pattern mining with formal concept analysis. In *Proc. the 22nd International Joint Conference on Artificial Intelligence*, 2011.

[18] Jaccard P. The distribution of the flora in the alpine zone. *New Phytologist*, 1912, 11: pp.37-50.

[19] Tan P N, Steinbach M, Kumar V. In *Introduction to Data Mining*, Addison-Wesley

Longman Publishing Co., Inc. Boston, MA, USA, 2005.

[20] Wang L, Wai-Lok Cheung D, Cheng R, Lee S D, Yang X S. Efficient mining of frequent item sets on large uncertain databases. *IEEE Transactions on Knowledge & Data Engineering*, 2012, 24(12): pp.2170-2183.

[21] Tong Y X, Chen L, Cheng Y, Yu P S. Mining frequent itemsets over uncertain databases. In *Proc. the VLDB Endowment*, 2012, 5(11): pp.1650-1661.

[22] Sandeep K S, Ganesh W, Niresh S. A review: data mining with fuzzy association rule mining. *International Journal of Engineering Research & Technology (IJERT)*, 2012, 1(5).

[23] Prabha K S, Lawrance R. Mining fuzzy frequent itemset using compact frequent pattern (CFP) tree algorithm. In *Proc. the International Conference on Computing and Control Engineering*, 2012.

[24] C Johnson S. Hierarchical clustering schemes. *Psychometrika*, 1967, 32(3): pp.241-254.

[25] Karypis G, (Sam) Han E H, Kumar V. Chameleon: a hierarchical clustering algorithm using dynamic modeling. *IEEE Computer*, 1999, 32(8): pp.68-75.

[26] Corral A, Manolopoulos Y, Theodoridis Y, Vassilakopoulos M. Algorithms for processing K-closest-pair queries in spatial databases. *Data & Knowledge Engineering*, 2004, 49 (1): pp.67-104.

[27] Guttman A. R-trees: a dynamic index structure for spatial searching. In *Proc. ACM SIGMOD Conference*, 1984, pp.47-57.

[28] Ibaraki T. *Annals of Operations Research*, Scientific Publishing Company, 1987.

[29] Xiao C, Wang W, Lin X M, Xu Y J, Wang G R. Efficient similarity joins for near-duplicate detection. *ACM Transactions on Database Systems (TODS)*, 2011, 36(3).

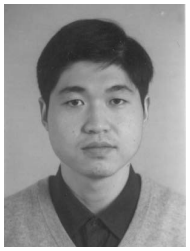
[30] Sun T Y, Shu C C, Li F, Yu H Y, Ma L L, Fang Y T. An efficient hierarchical clustering method for large datasets with map-reduce. In *Proc. the International Conference on Parallel and Distributed Computing, Applications and Technologies*, 2009.

[31] Bruynooghe M. Recent results in hierarchical clustering: I-the reducible neighborhoods clustering algorithm. *Int. J. Patt. Recogn. Artif. Intell.*, 1993, 7(3): pp.541-571.

[32] Siegfried K. Multivariate tests based on pairwise distance or similarity measures. In *Proc. the 6th Conference on Multivariate Distributions with Fixed Marginals*, Tartu, Estonia, June 2007.

Xin Xu received her Ph.D degree in computer science in School of Computing from National University of Singapore in Singapore in 2006. She is currently a Senior Research Engineer in Science and Technology on Information System Engineering Laboratory in Nanjing Research Institute of Electronic Engineering in Nanjing, China. Her research interests are in the area of artificial intelligence, data mining and pattern recognition.

Jiaheng Lu received his Ph.D degree in computer science in School of Computing from National University of Singapore in Singapore in 2006. He is currently an associated professor in Department of Computer Science of University of Helsinki, Finland. His research interests include multi-model database management systems, semantic string processing and job optimization for big data platform.



Wei Wang received his Ph.D degree in electrical and computer engineering from National University of Singapore in Singapore in 2008. He is currently an associated professor in Department of Computer Science and Technology of Nanjing University in Nanjing, China. His research interests are in the area of wireless sensor networks and pattern recognition.